EXHIBIT A

FUNDAMENTALS OF SPEECH RECOGNITION

Lawrence Rabiner
Biing-Hwang Juang

PTR Prentice Hall
Englewood Cliffs, New Jersey 07632

Library of Congress Cataloging-in-Publication Data

Rabiner, Lawrence R., 1943–
Fundamentals of speech recognition / Lawrence Rabiner, Biing-Hwang Juang.
p. cm.
Includes bibliographical references and index.
ISBN 0–13–015157–2
1. Automatic speech recognition. 2. Speech processing systems.
I. Juang, B. H. (Biing-Hwang) II. Title.
TK7895.S65R33 1993
006.4'54—dc20 92–34093
CIP

Editorial production and interior design: *bookworks*
Acquisitions Editor: *Karen Gettman*
Cover Designer: *Ben Santora*
Manufacturing Buyer: *Mary Elizabeth McCartney*

Published by PTR Prentice-Hall, Inc.
A Simon & Schuster Company
Englewood Cliffs, New Jersey 07632

The publisher offers discounts on this book when ordered in bulk quantities. For more information, contact:

Corporate Sales Department
PTR Prentice Hall
113 Sylvan Avenue
Englewood Cliffs, NJ 07632

Phone: 201-592-2863
FAX: 201-592-2249

Printed in the United States of America
10 9 8 7 6 5 4 3 2 1

ISBN 0-13-015157-2

Prentice-Hall International (UK) Limited, *London*
Prentice-Hall of Australia Pty. Limited, *Sydney*
Prentice-Hall Canada Inc., *Toronto*
Prentice-Hall Hispanoamericana, S.A., *Mexico*
Prentice-Hall of India Private Limited, *New Delhi*
Prentice-Hall of Japan, Inc., *Tokyo*
Simon & Schuster Asia Pte. Ltd., *Singapore*
Editora Prentice-Hall do Brasil, Ltda., *Rio de Janeiro*

Chapter 8

LARGE VOCABULARY CONTINUOUS SPEECH RECOGNITION

8.1 INTRODUCTION

Throughout this book we have developed a wide range of tools, techniques, and algorithms for attacking several fundamental problems in speech recognition. In the previous chapter we saw how the different techniques came together to solve the connected word recognition problem. In this chapter we extend the concepts to include issues needed to solve the large vocabulary, continuous speech recognition problem. We will see that the fundamental ideas need modification because of the use of subword speech units; however, a great deal of the formalism for recognition, based on word units, is still preserved.

The standard approach to large vocabulary continuous speech recognition is to assume a simple probabilistic model of speech production whereby a specified word sequence, W, produces an acoustic observation sequence Y, with probability $P(W, Y)$. The goal is then to decode the word string, based on the acoustic observation sequence, so that the decoded string has the maximum a posteriori (MAP) probability, i.e.,

$$\hat{W} \ni P(\hat{W}|Y) = \max_{W} P(W|Y). \tag{8.1}$$

Using Bayes' Rule, Equation (8.1) can be written as

$$P(W|Y) = \frac{P(Y|W)P(W)}{P(Y)}. \tag{8.2}$$

Since $P(Y)$ is independent of W, the MAP decoding rule of Eq. (8.1) is

$$\hat{W} = \arg\max_{W} P(Y|W)P(W). \tag{8.3}$$

The first term in Eq. (8.3), $P(Y|W)$, is generally called the acoustic model, as it estimates the probability of a sequence of acoustic observations, conditioned on the word string. The way in which we compute $P(Y|W)$, for large vocabulary speech recognition, is to build statistical models for subword speech units, build up word models from these subword speech unit models (using a lexicon to describe the composition of words), and then postulate word sequences and evaluate the acoustic model probabilities via standard concatenation methods. Such methods are discussed in Sections 8.2–8.4 of this chapter.

The second term in Eq. (8.3), $P(W)$, is generally called the language model, as it describes the probability associated with a postulated sequence of words. Such language models can incorporate both syntactic and semantic constraints of the language and the recognition task. Often, when only syntactic constraints are used, the language model is called a grammar and may be of the form of a formal parser and syntax analyzer, an N-gram word model ($N = 2, 3, \ldots$), or a word pair grammar of some type. Generally such language models are represented in a finite state network so as to be integrated into the acoustic model in a straightforward manner. We discuss language models further in Section 8.5 of this chapter.

We begin the chapter with a discussion of subword speech units. We formally define subword units and discuss their relative advantages (and disadvantages) as compared to whole-word models. We next show how we use standard statistical modeling techniques (i.e., hidden Markov models) to model subword units based on either discrete or continuous densities. We then show how such units can be trained automatically from continuous speech, without the need for a bootstrap model of each of the subword units. Next we discuss the problem of creating and implementing word lexicons (dictionaries) for use in both training and recognition phases. To evaluate the ideas discussed in this chapter we use a specified database access task, called the DARPA Resource Management (RM) task, in which there is a word vocabulary of 991 words (plus a silence or background word), and any one of several word grammars can be used. Using such a system, we show how a basic set of subword units performs on this task. Several directions for creating subword units which are more specialized are described, and several of these techniques are evaluated on the RM task. Finally we conclude the chapter with a discussion of how task semantics can be applied to further constrain the recognizer and improve overall performance.

8.2 SUBWORD SPEECH UNITS

We began Chapter 2 with a discussion of the basic phonetic units of language and discussed the acoustic properties of the phonemes in different speech contexts. We then argued that the acoustic variability of the phonemes due to context was sufficiently large and not well understood, that such units would not be useful as the basis for speech models for recognition. Instead, we have used whole-word models as the basic speech unit, both for

isolated word recognition systems and for connected word recognition systems, because whole words have the property that their acoustic representation is well defined, and the acoustic variability occurs mainly in the region of the beginning and the end of the word. Another advantage of using whole-word speech models is that it obviates the need for a word lexicon, thereby making the recognition structure inherently simple.

The disadvantages of using whole-word speech models for continuous speech recognition are twofold. First, to obtain reliable whole-word models, the number of word utterances in the training set needs to be sufficiently large, i.e., each word in the vocabulary should appear in each possible phonetic context several times in the training set. In this way the acoustic variability at the beginning and at the end of each word can be modeled appropriately. For word vocabularies like the digits, we know that each digit can be preceded and followed by every other digit; hence for an 11-digit vocabulary (zero to nine plus oh), there are exactly 121 phonetic contexts (some of which are essentially identical). Thus with a training set of several thousand digit strings, it is both realistic and practical to see every digit in every phonetic context several times. Now consider a vocabulary of 1000 words with an average of 100 phonetic contexts for both the beginning and end of each word. To see each word in each phonetic context exactly once requires $100 \times 1000 \times 100 = 10$ million carefully designed sentences. To see each combination 10 times requires 100 million such sentences. Clearly, the recording and processing of such homogeneous amounts of speech data is both impractical and unthinkable. Second, with a large vocabulary the phonetic content of the individual words will inevitably overlap. Thus storing and comparing whole-word patterns would be unduly redundant because the constituent sounds of individual words are treated independently, regardless of their identifiable similarities. Hence some more efficient speech representation is required for such large vocabulary systems. This is essentially the reason we use subword speech units.

There are several possible choices for subword units that can be used to model speech. These include the following:

- **Phonelike units** (PLUs) in which we use the basic phoneme set (or some appropriately modified set) of sounds but recognize that the acoustic properties of these units could be considerably different than the acoustic properties of the "basic" phonemes [1–7]. This is because we define the units based on linguistic similarity but model the unit based on acoustic similarity. In cases in which the acoustic and phonetic similarities are roughly the same (e.g., stressed vowels) then the phoneme and the PLU will be essentially identical. In other cases there can be large differences and a simple one-to-one correspondence may be inadequate in terms of modeling accuracy. Typically there are about 50 PLUs for English.
- **Syllable-like units** in which we again use the linguistic definition of a syllable (namely a vowel nucleus plus the optional initial and final consonants or consonant clusters) to initially define these units, and then model the unit based on acoustic similarity. In English there are approximately 10,000 syllables.
- **Dyad or demisyllable-like units** consisting of either the initial (optional) consonant cluster and some part of the vowel nucleus, or the remaining part of the vowel nucleus and the final (optional) consonant cluster [8]. For English there is something on the

order of 2000 demisyllable-like units.

- **Acoustic units**, which are defined on the basis of clustering speech segments from a segmentation of fluent, unlabeled speech using a specified objective criterion (e.g., maximum likelihood) [9]. Literally a codebook of speech units is created whose interpretation, in terms of classical linguistic units, is at best vague and at worst totally nonexistent. It has been shown that a set of 256–512 acoustic units is appropriate for modeling a wide range of speech vocabularies.

Consider the English word *segmentation*. Its representation according to each of the above subword unit sets is

- **PLUs**: /s/ /ɛ/ /g/ /m/ /ə/ /n/ /t/ /e^{y}/ /sh/ /ə/ /n/ (11 units)
- **syllables**: /seg/ /men/ /ta/ /tion/ (4 syllables)
- **demisyllables**: /sɛ/ /ɛg/ /mə/ /ən/ /te^{y}/ /e^{y}sh/ /shə/ /ən/ (8 demisyllables)
- **acoustic units**: 17 111 37 3 241 121 99 171 37 (9 acoustic units).

We see, from the above example, that the number of subword units for this word can be as small as 4 (from a set of 10,000 units) or as large as 11 (from a set of 50 units).

Since each of the above subword unit sets is capable of representing any word in the English language, the issues in the choice of subword unit sets are the context sensitivity and the ease of training the unit from fluent speech. (In addition, for acoustic units, an issue is the creation of a word lexicon since the units themselves have no inherent linguistic interpretation.) It should be clear that there is no ideal (perfect) set of subword units. The PLU set is extremely context sensitive because each unit is potentially affected by its predecessors (one or more) and its followers. However, there is only a small number of PLUs and they are relatively easy to train. On the other extreme are the syllables which are longest units and are the least context sensitive. However, there are so many of them that they are almost as difficult to train as whole-word models.

For simplicity we will initially assume that we use PLUs as the basic speech units. In particular we use the set of 47 PLUs shown in Table 8.1 (which includes an explicit symbol for silence –h#). For each PLU we show an orthographic symbol (e.g., aa) and a word associated with the symbol (e.g., father). (These symbols are essentially identical to the ARPAPET alphabet of Table 2.1; lowercase symbols are used throughout this chapter for consistency with the DARPA community.) Table 8.2 shows typical pronunciations for several words from the DARPA RM task in terms of the PLUs in Table 8.1. A strong advantage of using PLUs is the ease of creating word lexicons of the type shown in Table 8.2 from standard (electronic) dictionaries. We will see later in this chapter how we exploit the advantages of PLUs, while reducing the context dependencies, by going to more specialized PLUs which take into consideration either the left or right (or both) contexts in which the PLU appears.

One problem with word lexicons of the type shown in Table 8.2 is that they don't easily account for variations in word pronunciation across different dialects and in the context of a sentence. Hence a simple word like "a" is often pronounced as /ey/ in isolation (e.g., the